Data Source & Corpus Building

Embeddings are vectors - numerical representations of semantic content in a higher dimensional geometry.

Master Corpus

0.0.A

Lexical Corpus

0.0.B

Analysis Corpus

Shadow Corpus

Shadow Subsample

Semantic Classification

Embeddings are vectors - numerical representations of semantic content in a higher dimensional geometry.

WORD EMBEDDINGS

LLM

OPEN AI Platform

VARIABLES:
model
prompt
temperature
inference explicitly

EVALUATE

OPTIMIZE

score < 0.8

score < 0.8

Embeddings

Embeddings are vectors - numerical representations of semantic content in a higher dimensional geometry.

Dimension Reduction

OUTPUT LAYERS

Last
Mid 4 mean
Weighted
Last 4 mean
All mean

DIMENSIONS

25 , 50, 100, 200

ALGORITHM

PCA, UMAP, GRP, AE

kNN recall

Drift corr.

Trust

Evaluation

Run 1

- "A"
- "A", "B"
- "C"

Run 2

- "A"
- "A", "C"
- "C"

Exact match
Exact match kappa
Jaccard metric

20%

VALIDATE

80%

TRAINING

0.2.B

Shadow Subsample

SCIBERT

CLUSTERING

Evaluation

F1 <

F1 >

0.2.C

Machine Learning

Embeddings are vectors - numerical representations of semantic content in a higher dimensional geometry.

10%

TEST

100%

FINAL

Significance

Analysis Coprus:
Geometric semantic properties from context embeddings & referential and explanatory classifications from machine learning inference

Drift Metrics
On Embeddings

Lexical Analysis

Analysis

Embeddings are vectors - numerical representations of semantic content in a higher dimensional geometry.

Process Diagram

Here are the methodological map of all the conceptual and logical steps of the process.

Legend

The X axis shows data separation: some of the corpus is contained to train the ML classifier for example, so it trains on the same kind of data, but no exact overlapping data.

The Y-axis shows derivative steps from the root data-source.

The Z-axis is the order of operations, or the sequence, in which the scripts and computations are executed.

